

Analisi Predittiva

CT0429

aa 2022/23

Primo appello

Con Soluzioni

Gennaio, 2023

Cognome: _____ Nome: _____

Matricola: _____ Firma: _____

ISTRUZIONI (DA LEGGERE ATTENTAMENTE).

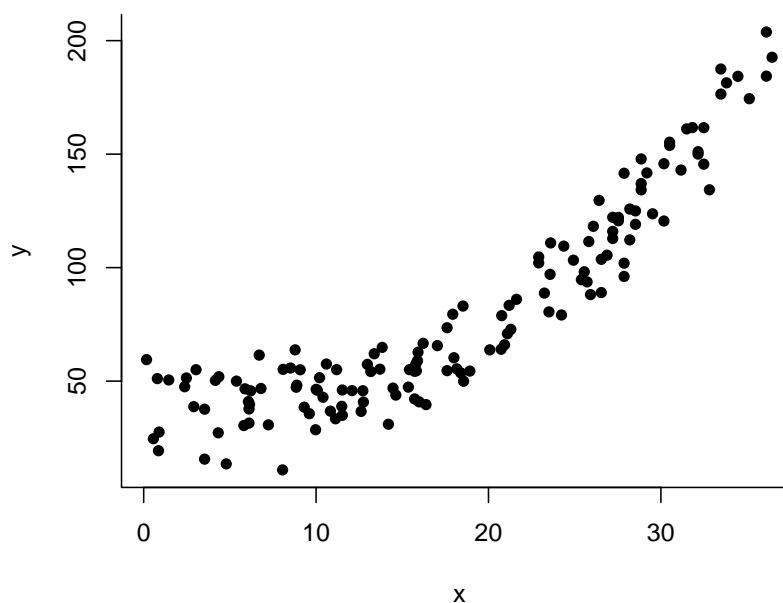
Assicuratevi di aver scritto nome cognome e matricola sia qui che sul file Rmarkdown disponibile su Moodle. Il tempo a disposizione per completare tutto l'esame (la parte scritta e la parte su Moodle) è di **90 minuti**.

Nessuno studente può lasciare l'aula fino a che la docente non avrà verificato che tutti abbiano consegnato sia il compito scritto che il file Rmarkdown. Dopo la consegna attendete che la docente dia il permesso di lasciare l'aula.

Question:	1	2	3	4	Total
Points:	5	3	5	5	18
Score:					

Question 1 (5 points)

Il dataset `df` contiene misure per un campione di 145 osservazioni di due variabili continue (`x` e `y`) mostrate nel grafico sottostante:



Queste vengono usate per stimare il seguente modello:

```
fit <- lm(y~x, data = df)
as.numeric(c(coef(fit), summary(fit)$sigma))
```

```
[1] 6.5 4.0 20.0
```

Come da codice sovrastante, si ottengono le seguenti stime: $(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}) = (6.5, 4, 20)$

Si crea quindi un vettore `e` che contiene i 145 residui del modello, i.e. $e_i = (y_i - \hat{y}_i)$

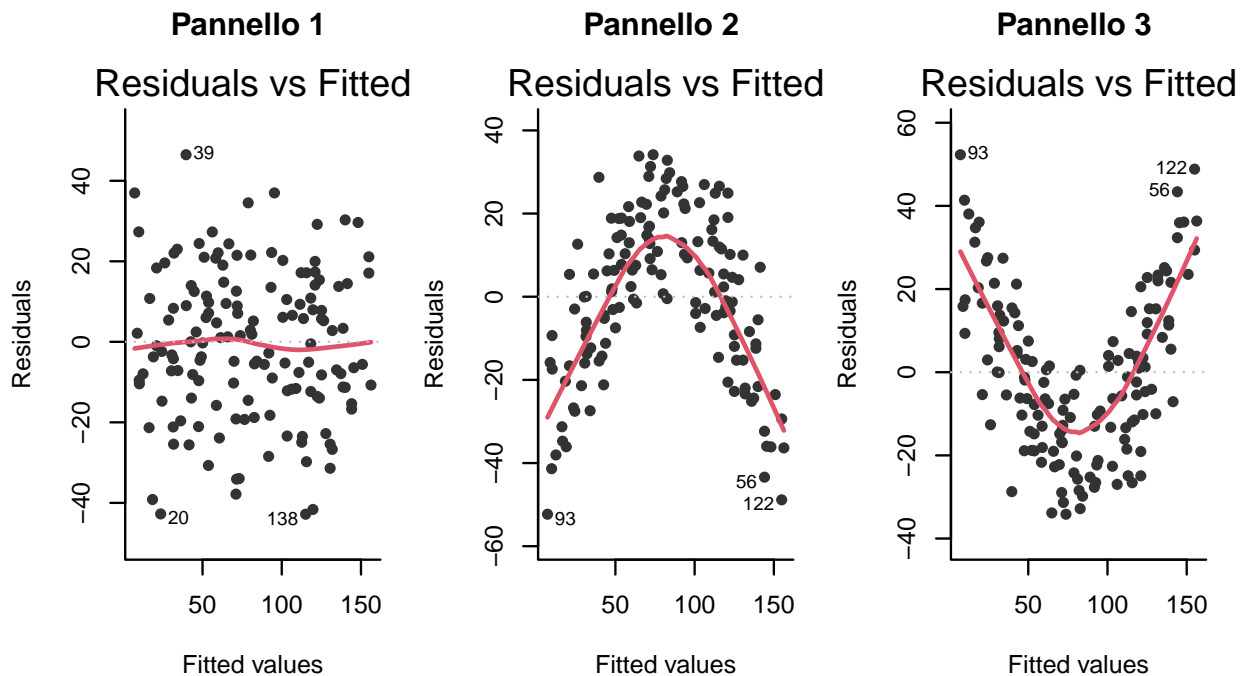
```
e <- residuals(fit)
```

i) Qual è il valore di $\sum_{i=1}^n e_i$ (i.e. `sum(e)`)?

ii) Qual è il valore di $\sum_{i=1}^n e_i^2$ (`sum(e^2)`)?

iii) Sapendo che $\bar{y} = 7$, è possibile determinare il valore di \bar{x} ? Se sì indicarne il valore:

- iv) Quale dei Pannelli della figura sottostante è probabile mostri il grafico dei residui del modello fit? Si motivi la risposta. Pannello _____ perché



Solution:

i) Proprietà dei residui: $\sum_{i=1}^n e_i = 0$

ii) Dato che $\hat{\sigma}^2 = \sum_{i=1}^n e_i^2 / (n - 2)$ si ha che

$$\sum_{i=1}^n e_i^2 = \hat{\sigma}^2 * (n - 2) = 400 * 143 = 57200$$

iii) Si ha che $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$. Cambiando l'ordine dell'equazione si ha:

$$\bar{x} = (\bar{y} - \hat{\beta}_0) / \hat{\beta}_1 = (7 - 6.5) / 4 = 0.125$$

iv) Il pannello 3 dato che si nota nel grafico di dispersione tra X e Y che la relazione tra le due variabili è quadratica (convessa) e ci si può aspettare di trovare la stessa relazione tra valori stimati (che sono una trasformazione lineare di x) e residui.

Question 2 (3 points)

Un'azienda desidera monitorare la relazione tra il numero di clienti gestito da un addetto alle vendite e il volume di vendite generato dall'addetto. Per un campione di 45 addetti vengono raccolte le seguenti informazioni:

- **nAccounts**: il numero di clienti gestito da un addetto alle vendite
- **filiale**: la filiale presso cui presta servizio l'addetto
- **volume**: il volume totale di vendite legato all'addetto

Vengono stimati due modelli:

```
fit1 <- lm(volume ~ nAccounts+filiale, data = df)
fit2 <- lm(volume ~ nAccounts*filiale, data = df)
coef(fit1)
```

(Intercept)	nAccounts	filialeB
294.9	1.6	-28.5

```
coef(fit2)
```

(Intercept)	nAccounts	filialeB	nAccounts:filialeB
295.416	1.548	-29.358	0.021

Il modello `fit1` viene poi usato per stimare il volume di due addetti:

```
nd
```

	nAccounts	filiale
Luigi	30	A
Chiara	30	B

```
predict(fit1, newdata = nd)
```

```
[1] NA NA
```

```
dim(df)
```

```
[1] 45 3
```

- i) Per quale delle due osservazioni risulta più alta la stima della variabile **volume** ottenuta con il modello `fit1`? Osservazione _____ perché

- ii) Quale dei modelli stimati permette di verificare se l'effetto di **nAccounts** su **volume** è lo stesso nelle due filiali?

Solution:

- (i) La stima risulterà più alta per Luigi: i due responsabili hanno lo valore per **nAccounts**, ma l'effetto di essere nella filiale B è negativo, quindi a parità di **nAccounts** la stima per persone nella filiale B è più bassa di 28.5 unità.
- (ii) Il modello **fit2** stima un parametro aggiuntivo che descrive il differenziale dell'effetto di **nAccounts** su **volume** nelle due filiali: verificare se questo parametro è pari a 0 permette di verificare se l'effetto **nAccounts** è lo stesso nelle due filiali

Question 3 (5 points)

Un'azienda desidera monitorare la relazione tra il numero di clienti gestito da un addetto alle vendite e il margine premiale pagato all'addetto. Per un campione di 73 addetti vengono derivate le seguenti informazioni

- **nAccounts**: il numero di clienti gestito da un addetto alle vendite
- **payBenefit**: la premialità pagata all'addetto nel mese precedente

Alcune statistiche descrittive per il dataset sono riportate qui sotto:

nAccounts	payBenefit.V1
Min. :10	Min. :13
1st Qu.:31	1st Qu.:49
Median :44	Median :66
Mean :46	Mean :60
3rd Qu.:62	3rd Qu.:71
Max. :80	Max. :93

Vengono stimati i seguenti modelli

```
fit1 <- lm(payBenefit ~ nAccounts, data = df)
fit_Xt <- lm(payBenefit ~ I(nAccounts-10), data = df)
fit_lxly <- lm(log(payBenefit) ~ log(nAccounts), data = df)
```

Si derivano poi caratteristiche della stima del modello `fit1`:

```
summary(fit1)
```

...

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	33.1449	3.4208	9.69	1.3e-14
nAccounts	0.5938	0.0683	8.70	8.4e-13

Residual standard error: 12 on 71 degrees of freedom

Multiple R-squared: 0.516, Adjusted R-squared: 0.509

F-statistic: 75.7 on 1 and 71 DF, p-value: 8.43e-13

Infine il modello `fit1` viene usato per predire i valori medi di `payBenefit` per due addetti alle vendite. Oltre alla stima puntuale viene calcolato un intervallo di confidenza al 98%:

```
nd
```

	nAccounts
Marco	45
Sara	90

```
predict(fit1, newdata = nd, interval = "confidence", level = .98)
```

```
[1] NA NA
```

- i) Con le informazioni a vostra disposizione, è possibile sapere quale è la stima del valore del coefficiente angolare β_1 che descrive l'effetto di `I(nAccounts-10)` nel modello `fit_Xt`? Se sì, se ne indichi il valore.
- ii) Con le informazioni a vostra disposizione, è possibile sapere quale è il valore di R^2 nel modello `fit_Xt`? Se sì, se ne indichi il valore.
- iii) Con le informazioni a vostra disposizione, è possibile sapere quale è la stima del valore del coefficiente angolare β_1 che descrive l'effetto di `log(nAccounts)` nel modello `fit_lXlY`? Se sì, se ne indichi il valore.
- iv) Con le informazioni a vostra disposizione, è possibile sapere quale è il valore di R^2 nel modello `fit_lXlY`? Se sì, se ne indichi il valore.
- v) Si indichi per quale dei due dipendenti l'intervallo di confidenza della stima di `payBenefit` è più ampio:

Solution: Il modello `fit_Xt` usa la stessa variabile risposta mentre la variabile esplicativa viene inserita con una trasformazione lineare. Se il modello `fit` è definito come:

$$y_i = \beta_0 + \beta_1 x + \epsilon_i$$

e il modello `fit_Xt` è definito come:

$$y_i = \gamma_0 + \gamma_1(x - 10) + \epsilon_i = (\gamma_0 - 10 * \gamma_1) + \gamma_1 x + \epsilon_i$$

se ne deduce che $\gamma_0 - 10 * \gamma_1 = \beta_0$ e $\gamma_1 = \beta_1$. Dato che i due modelli differiscono solo nel valore numerico dell'intercetta spiegano la stessa variabilità di Y e hanno quindi lo stesso valore di R^2 . Questo non è più vero quando viene applicata una trasformazione non lineare sia a variabile esplicativa che a alla risposta.

- (i) 0.5938
- (ii) 0.516
- (iii) Non è possibile
- (iv) Non è possibile
- (v) L'incertezza della stima è sempre più larga per valori della variabile esplicativa che sono più distanti dalla media: vi è più incertezza nella stima derivata per Sara.

Question 4 (5 points)

Un'azienda che gestisce un sito di vendite di oggetti di seconda mano desidera identificare quali siano i fattori che fanno aumentare la probabilità che un annuncio venga chiuso perché è avvenuta una vendita tramite il sito. Ad un campione di 40 utenti che hanno chiuso un annuncio viene chiesto se la chiusura dell'annuncio è dovuta ad una vendita tramite il sito e vengono raccolte alcune informazioni relative ad utente ed annuncio. In particolare si hanno le seguenti informazioni:

- **vendita**: una variabile dicotomica che indica se la chiusura di un annuncio è dovuta ad una vendita tramite il sito
- **Nfoto**: una variabile categoriale che indica il numero di foto allegate all'annuncio. La variabile può avere valori: 0, 1-3, 4-6
- **DiffPrezzoMedio**: una variabile continua che indica la differenza relativa tra il prezzo richiesto nell'annuncio e il prezzo medio per oggetti nella stessa categoria

Vengono stimati due modelli alternativi:

```
fit1 <- glm(vendita ~ DiffPrezzoMedio, family = "binomial", data = df)
fit2 <- glm(vendita ~ Nfoto+DiffPrezzoMedio, family = "binomial", data = df)

summary(fit1)$coef
```

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.344	0.394	0.87	0.383
DiffPrezzoMedio	-0.073	0.033	-2.22	0.026

Usando il modello `fit1` si deriva la funzione che stima la probabilità di chiudere un annuncio con una vendita in funzione della differenza relativa del prezzo:

```
nd <- data.frame(DiffPrezzoMedio = seq(-20,26,by = 0.25))
plot(nd$DiffPrezzoMedio,predict(fit1, newdata = nd,type = "response"),type = "l")
```

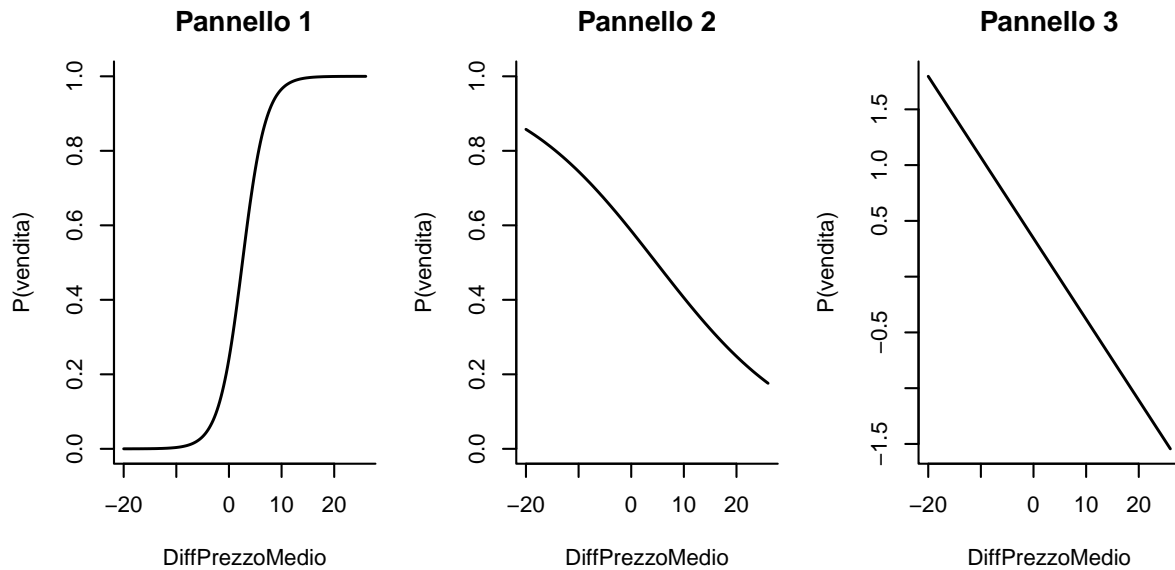
Successivamente si confronta la bontà di adattamento dei due modelli utilizzando un test del rapporto di verosimiglianza (likelihood ratio test) tramite la funzione `anova`:

```
anova(fit1, fit2, test = "LRT")
```

Analysis of Deviance Table

```
Model 1: vendita ~ DiffPrezzoMedio
Model 2: vendita ~ Nfoto + DiffPrezzoMedio
```

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	38	49.5			
2	36	46.4	2	3.08	0.21



- (i) In quale dei pannelli nella figura sovrastante è più credibile che sia mostrata la relazione stimata dal modello `fit1`? Si motivi la risposta. Pannello _____ perché
- (ii) Si scriva in forma esplicita il modello specificato con `fit2`
- (iii) Si specifichino ipotesi nulla ed alternativa del test del rapporto di verosimiglianza derivato dalla funzione `anova` dando indicazione su come interpretare l'output della funzione che confronta i modelli `fit1` e `fit2`

Solution: i) Pannello 2. La stima del coefficiente di regressione è negativa e poco significativa/prossima allo 0: all'aumentare della differenza di prezzo vi è una debole diminuzione della probabilità di vendita. Gli altri pannelli mostrano una forte relazione positiva (Pannello 1) o una relazione negativa ma su una scala che non può essere quella della probabilità (valori negativi sull'asse delle Y).

ii)

$$Vendita_i \sim \text{Bern}(p(\text{DiffPrezzoMedio}_i, \text{Nfoto}_i))$$

dove

$$\text{logit}(p(\text{DiffPrezzoMedio}_i, \text{Nfoto}_i)) = \beta_0 + \beta_1 \text{DiffPrezzoMedio}_i + \beta_2 \text{Nfoto}(1-3)_i + \beta_3 \text{Nfoto}(4-6)_i$$

iii)

$$H_0 : \beta_2 = \beta_3 = 0 \quad VS \quad H_1 : \beta_2 \text{ or } \beta_3 \neq 0$$

Il p-value è abbastanza grande: non possiamo rifiutare l'ipotesi nulla. Il numero di foto non sembra avere un forte impatto sulla probabilità che un annuncio si concluda con una vendita.